MATHÉMATIQUES ET SCIENCES HUMAINES

S. RÉGNIER

Sur quelques aspects mathématiques des problèmes de classification automatique

Mathématiques et sciences humaines, tome 82 (1983), p. 13-29

<http://www.numdam.org/item?id=MSH_1983__82__13_0>

Math. Sci. hum. (21^{e} année, n°82, *1983*, p.13-29)

SUR QUELQUES ASPECTS MATHEMATIQUES DES PROBLEMES DE CLASSIFICATION AUTOMATIQUE*

S. REGNIER

Maison des Sciences de l'Homme, Centre de Calcul

1. ASPECT ALGEBRIQUE

La classification automatique a pour but de définir sur un ensemble d'objets deux à deux comparables, une partition (en groupes disjoints et complémentaires) qui respecte au mieux les ressemblances entre objets. Il convient que la ressemblance de deux objets soit grande lorsqu'ils figurent dans le même groupe et petite dans le cas contraire.

1.1. *Mesure de la Ressemblance: Distances et similarités*

Le choix d'une mesure de ressemblance dépend essentiellement de la nature des objets étudiés.

En général, chacun est décrit par un certain nombre de critères a_h correspondant à divers "points de vue" et l'ensemble des a_h peut se résumer par une application:

$$E \xrightarrow{a} \begin{cases} F_1 x F_2 \ldots F_h \ldots x F_p \\ F_h = \text{ensemble des valeurs du critière } a_h \end{cases}$$

ensemble d'objets

où

$$a = (a_1, a_2, \ldots a_h, \ldots a_p)$$

Distance

Lorsque les ensembles F_h possèdent une métrique naturelle (par exemple lorsque les valeurs de a_h se représentent naturellement par des nombres réels), on peut poser

$$d_{ij}^h = \text{Distance entre } a_h(i) \text{ et } a_h(j) \text{ dans } F_h$$
$$i \text{ et } j \in E$$

* ICC Bulletin. 1965. Vol. 4, pp. 175–191.

et définir une distance globale d_{ij} entre les objets i et j comme la moyenne des p distances d_{ij}^h, associées aux divers critères. On peut d'ailleurs pondérer les critères selon leur importance et poser

$$d_{ij} = \sum_{h=1}^{p} p_h d_{ij}^h \begin{cases} p_h > 0 \\ \sum p_h = 1 \end{cases}$$

autrement dit, d_{ij} = moyenne des d_{ij}^h pondérés par les p_h. On notera que cette distance est nulle si, et seulement si

$$a_h(i) = a_h(j) \; \forall h$$

c'est à dire si les deux objects sont identiques à tous points de vues; ce qui ne signifie pas qu'ils soient totalement identiques.

Similarité

Lorsque les ensembles F_h n'ont aucune métrique naturelle, mais sont finis, chaque critère a_h définit par lui même une partition de l'ensemble d'objets E, ou chaque groupe est formé des objets identiques selon le critère a_h:

$$i \text{ et } j \text{ dans le même groupe} \Leftrightarrow a_h(i) = a_h(j)$$

Cette relation d'équivalence entre i et j peut être représentée par une variable logique:

$$\begin{cases} s_{ij}^h = 1 \text{ si } a_h(i) = a_h(j) \\ \quad = 0 \text{ sinon} \end{cases}$$

Dans ce cas la moyenne:

$$s_{ij} = \frac{1}{p} \sum_{1}^{p} s_{ij}^h$$

représente la proportion de critères a_h qui prennent la même valeur sur sur les objets i et j. Ici encore on peut pondérer les critères et définir la *similarité moyenne entre i et j.*

$$s_h = \sum p_h s_{ij}^h \begin{cases} p_h > 0 \\ \sum p_h = 1 \end{cases}$$

L'intérêt de cette définition apparaîtra au §2.2.

Lien entre similarité et distance

On peut définir dans F_h une distance:

$$d(f_1, f_2) = 0 \text{ si } f_1 = f_2$$
$$1 \text{ si } f_1 \neq f_2$$

à partir de là, les deux définitions ci-dessus sont étroitement liées par les relations

$$s_{ij}^h = 1 - d_{ij}^h$$
$$s_{ij} = 1 - d_{ij}$$

Ce qui suggère deux généralisations:

Définitions générales de la similarité de deux objets

1) Si l'ensemble de critères

$$E \xrightarrow{a} F$$

définit une distance d_{ij} entre les objets i et j on posera

$$s_{ij} = 1 - d_{ij}$$

2) plus généralement une similarité est une fonction

$$E \times E \xrightarrow{s} R$$

qui vérifie les axiomes

— $S_1 : s_{ij} \leqslant 1 \qquad s_{ii} = 1$
— $S_2 : s_{ij} = s_{ji}$
— S_3: la similarité est d'autant plus forte que les objets sont plus ressemblants, et vaut 1 seulement si les objets sont "à tous points de vue identiques", et $s_{ij} = 1$ entraine pour tout k $s_{ik} = s_{jk}$.

1.2. *Structure de l'ensemble des partitions de E*

Soit n le cardinal de E, le nombre d'objets à classer.

Notre but est de définir une "partition" de E: un ensemble de parties disjointes et complémentaires.

La relation "i et j sont dans la même partie" va être une relation d'équivalence, et chaque partie une classe d'équivalence.

Considérons $n \times n$ variables logiques:

$$\begin{cases} g_{ij} = 1 \text{ si } i \text{ et } j \text{ sont dans le même groupe} \\ \quad = 0 \text{ sinon} \end{cases}$$

La fonction $i, j \xrightarrow{g} g_{ij}$ est alors la fonction caractéristique ou fonction indicatrice du graphe dans $E \times E$, de la relation considérée; g est une application de $E \times E$, sur l'ensemble à 2 éléments $\{0, 1\}$, dans l'ensemble R des nombres réels. C'est donc un élément de l'ensemble $R^{E \times E}$, qui est isomorphe à $R^{n \times n}$, espace vectoriel de dimension $n \times n$ sur le corps R.

P = ensemble des partitions, est une partie de

C = ensemble des relations sur $E \times E$

représenté dans $R^{n \times n}$ par le cube $(0,1)^{n \times n}$ des points g tel que

$$g_{ij} = 1 \text{ ou } 0 \ \forall i, j$$

$$P \subset (0,1)^{n \times n} \subset R^{n \times n}$$

La métrique de $R^{n \times n}$ définit une distance D telle que

$$D^2 = (g - g')^2 = \sum_{i \text{ et } j \in E} \sum (g_{ij} - g'_{ij})^2$$

Si g et g' sont deux relations, $g_{ij} = 1$ ou 0; D^2 représente le cardinal de la différence symétrique des parties correspondantes dans $E \times E$; et dans le cas de partitions:

D^2 = nombre de couples i, j qui sont classés ensemble par une seule des deux partitions.

L'intérêt de plonger P dans $R^{n \times n}$ tient aux propriétés des barycentres. Le barycentre des points g^h pondérés par les poids p_h est par définition

$$g = \sum p_h g^h$$

et vérifie $\forall X \in R^{n \times n}$

où l'on a désigné, si $A, B, \in R^{n \times n}$ par AB l'element $B - A \in R^{n \times n}$

$$F(X) = \sum p_h (Xg^h)^2 = (Xg)^2 + \sum p_h (gg^h)^2 = (Xg)^2 + F(g)$$

On voit que g réalise le minimum de $F(X)$ et que pour minimiser $F(X)$ sous une condition $X \in C$, il suffit de résoudre:

$$\begin{cases} \text{Min } (Xg)^2 \\ X \in C \end{cases}$$

1.3. *Partition Centrale*

Reprenons donc les similarités $s_{ij}^h = 1 - d_{ij}^h$ définies au paragraphe 1.1. La matrice S^h des s_{ij}^h pour chaque h est un élément de $R^{n \times n}$. La matrice s des similarités moyennes (s_{ij}) est le barycentre de ces vecteurs pondérés par les poids p_h.

Nous appellerons *partition centrale* tout vecteur $X \in P \subset R^{n \times n}$ qui réalise $(Xs)^2$ minimum; car elle réalise le minimum de $\sum p_h (Xs^h)^2$ (les vecteurs s^h sont d'ailleurs eux-mêmes des partitions lorsque s_{ij}^h indique $a_h(i) = a_h(j)$). X est en quelque sorte une projection de s sur l'ensemble des partitions dans $R^{n \times n}$, et se présente comme une solution naturelle du problème posé.

La fonction $(Xs)^2$ se simplifie si l'on remarque que:

$$\text{pour } X \in C, \text{ soit } X_{ij} = 0 \text{ ou } 1$$

X est à distance fixe du point H de coordonnées $\frac{1}{2}$

$$h_{ij} = \tfrac{1}{2} \forall\ i \text{ et } j$$
$$(XH)^2 = n^2/4$$

En termes géométriques le cube C est inscrit dans la sphère de centre H et de rayon $n^2/4$; par conséquent

$$(sX)^2 = (HX - Hs)^2 = (HX)^2 + (Hs)^2 - 2Hs \times HX$$
$$= n^2/4 + (Hs)^2 - 2Hs \times HX$$

Le point X cherché réalise le maximum du produit scalaire

$$Hs \times HX = \sum_i \sum_j (s_{ij} - \tfrac{1}{2})(x_{ij} - \tfrac{1}{2})$$

soit encore de la forme linéaire

$$\boxed{L(X) = \sum_{ji} t_{ij} x_{ij}}$$

où $$t_{ij} = s_{ij} - \tfrac{1}{2}$$

En introduisant une application a de E dans un ensemble F telle que

$$x_{ij} = 1 \Leftrightarrow a(i) = a(j)$$

par exemple une numérotation arbitraire des classes d'équivalence formant la partition X, on obtient une expression plus condensée

$$\boxed{L(X) = \sum_{f \in F} \sum_{a(i)=a(j)=f} t_{ji}}$$

$$L(f) = \sum_{a(i)=a(j)=f} t_{ij}$$

représente une sorte de compacité de la classe d'équivalence de numéro f, et $L(X)$ est la somme des compacités des différentes classes.

Notons qu'en maximisant

$$L(X) = \sum_{a(i)=a(j)} t_{ij}$$

On rend également maxima

$$M(X) = \sum t_{ij}(2x_{ij}-1) = \sum_{a(i)=a(j)} t_{ij} - \sum_{a(i)\neq a(j)} t_{ij}$$

Les valeurs de t_{ij} pour les objets classés différemment sont donc implicitement prise en compte Cette quantité $t_{ij}=s_{ij}-\frac{1}{2}$ peut-être appel ée "attraction des objets i et j".

On remarquera que le problème est encore invariant si les attractions t_{ij} sont toutes multipliées par une même constante positive. Dans $R^{n\times n}$ le point X qui réalise $HX \times Hs$ maximum, ne dépend que de la direction du vecteur Hs.

1.4. *Méthodes de Calcul des Partitions Centrales X*

1.4.1. *Programmation Linéaire*

Nous cherchons dans $R^{n\times n}$ un vecteur X représentatif d'une partition et qui réalise le maximum d'une forme linéaire $L(X)$

$$\begin{cases} \sum t_{ij}x_{ij} \text{ maximum} \\ (x_{ij}) = \text{partition de } E = \text{élément de } P \end{cases}$$

et cette dernière condition, en vertu des conventions déja faites, se décompose en

$-x_{ij} = 0$ ou 1 (relation)
$-x_{ih} \geqslant x_{ij}x_{ih}$ (transitive)
$-x_{ii} = 1$ (reflexive)
$-x_{ij} = x_{ji}$ (symétrique)

toutes ces conditions expriment que $x_{ij}=1$ définit entre les objets i et j une relation d'équivalence.

Les deux dernières permettent de n'envisager que les $\frac{n(n-1)}{2}$ inconnues:

$$x_{ij} \quad \text{pour} \quad 1 \leqslant i < j \leqslant n$$

et les contraintes suivantes suffisent à définir P

$$\left.\begin{matrix}\text{(a)}\ x_{ij} = 0 \text{ ou } 1 \\ \text{(b)}\ x_{ih} \geqslant x_{ij}x_{jh}\end{matrix}\right\} \quad i < h \text{ et } j \neq \text{de } i \text{ et } h.$$

P est naturellement un ensemble fini. Sa fermeture convexe est un polyèdre $C(P)$, dont le profil (ou ensemble des sommets) est P lui même, car il résulte de (a) qu'aucune relation n'est combinaison linéaire convexe d'autres relations. Le problème peut donc en principe être traité par un algorithme de programmation linéaire classique, ou de programmation en nombres entiers. Mais il faudrait tout d'abord écrire l'équation de $C(P)$ sous forme d'inéquations linéaires Il est à craindre qu'elles soient excessivement nombreuses puisque (b) définit déjà:

$$\frac{n(n-1)}{2}\frac{(n-2)}{} \quad \text{inéquations non linéaires.}$$

Plutôt que d'appliquer un algorithme universel, il semble naturel de tenir compte de la nature très particulière du polyèdre $C(P)$. On retiendra cependant que la solution X cherchée est général unique; et qu' exceptionnellement, $L(X)$ peut prendre sa valeur maxima sur toute une face du polyèdre $C(P)$, (HS est alors orthogonal à cette face). Les sommets de cette face constituent alors autant de partitions X optimales.

1.4.2. *Algorithme des Transferts*

A partir d'une partition X on peut construire une partition X' particulièrement voisine en effectuant le transfert d'un objet d'une classe

dans une autre. En représentant X par une numérotation abitraire des classes:

$$E \xrightarrow{a} \{1, 2, \ldots q\}$$

objet $i \longrightarrow$ N° de la classe qui contient i, on définit le transfert de l'objet K (de la classe L) dans la classe M, à partir de a:

$$\begin{array}{ll} T_k^m/a & \\ a \longrightarrow a' & \begin{cases} a'(K) = m \neq a(K) = 1 \\ a'(i) \;\; = a(i) \text{ si } i \neq k \end{cases} \end{array}$$

En répétant l'opération on peut parcourir tout l'ensemble P_q des partitions en q classes au plus (pour n objets $P_n = P$).

Il apparaît une nouvelle métrique T différente de la métrique Euclidienne: $T(X,X')$ = nombre minimum de transfertsnécessaires pour passer de X à X'. L'algorithme consiste à cheminer sur P_q par transfert, en choisissant chaque fois le transfert t_k^m/a qui augmente le plus la forme linéaire

$$L(X) = \sum t_{ij}\, x_{ij} = \sum_f \sum_{a(i)=a(j)=f} t_{ij}$$

Dans la représentation:

$$x_{ij} = \begin{array}{l} 1 \text{ si } a(i) = a(j) \\ 0 \text{ sinon,} \end{array}$$

le transfert ne modifie que la ligne et la colonne de l'objet k transferé:

$$\begin{array}{c} x_{jk}(= x_{kj}) \text{ passe de 1 à 0 si } a(j) = L \text{ et } j \neq k \\ \text{de 0 à 1 si } a(j) = M \end{array}$$

La variation de L est donc:

$$DL = 2 \sum_{a(j)=M} t_{jk} - 2 \sum_{a(j)=L \text{ et } j \neq k} t_{jk} \text{ et } j \neq k$$

On peut poser: $t_{kk} = 0 \;\forall k$

$$C_i^f = \sum_{a(j)=f} t_{ij} = \text{“attraction de l'objet } i \text{ par la classe } f\text{”}$$

Alors $DL = 2(C_k^M - C_k^L)$ donne le gain de transfert de l'objet k dans la classe M; et le choix du transfert le plus avantageux repose seulement sur la connaissance de la matrice des Ci^f (n lignes × q colonnes).

Au cours du transfert t_k^m/a, cette matrice se modifie selon des régles simples

$$\begin{cases} \text{les colonnes } f = L \text{ ou } M \text{ sont seules modifiées} \\ C_i^L \text{ devient } C_i^L - t_{ik} \\ C_i^M \text{ devient } C_i^M + t_{ik} \end{cases}$$

Règle d'arrêt

On peut arrêter le cheminement lorsqu'aucun transfert avantageux n'est possible:

$$\operatorname{Max}_{K,\,M} DL < 0$$

On obtient alors un maximum local de la fonction $L(X)$ au sens de la métrique T ci-dessus:

La partition X obtenue est meilleure que celles qu'on peut atteindre en effectuant un seul transfert à partir de X.

Il peut exister plusieurs maxima locaux.

L'algorithme considéré n'a donc pas les qualités habituelles d'un algorithme de programmation linéaire.

Il serait intéressant d'emprunter aux principes de la méthode du simplexe une procédure permettant de reconnaître si un maximum local est maximum général.

1.5. *Objets Equivalents (identiques a tous points de vue)*

Comme on le signalait en (1.1) il peut exister des couples d'objets i et j "identiques à tous points de vue" ils ont même représentation $a(i)$ dans l'ensemble $F = F_1 \times F_2 \ldots \times F_p$

$$a_h(i) = a_h(j) \; \forall h$$

On dira qu'ils sont équivalents.

Leur distance éventuelle est nulle (la distance sur F définit un "écart" sur E). Leur similarité moyenne est $s_{ij} = 1$ quels que soient les poids p_h utilisés. De plus $s_{ik} = s_{jk}$ quel que soit l'objet k. Il serait extrémement choquant que deux tels objets ne soient pas classés ensemble dans les partitions centrales X. L'algorithme des transferts permet d'établir une proposition un peu plus forte:

Proposition 1.5

Toute partition X qui est un maximum local est moins fine que la partition canonique en classe d'objets équivalents (l'intersection logique des partitions initiales).

En effet si deux objets j et k équivalents sont classés par X dans des classes différentes, on peut voir que l'un des deux transferts qui les réunit est strictement avantageux.

Les compacités t_{ik} et t_{ij} sont égales sauf pour $i=k$ ou j, puisque par convention d'écriture $t_{kk}=t_{jj}=0$, tandis que $t_{jk}=1/2$.

$$\text{Si } A(k) = L$$
$$A(j) = M$$

les 2 transferts envisagés offrent des gains doubles des quantités

$$G_1 = C_k^M - C_k^L \qquad G_2 = C_j^L - C_j^M$$

or $C_j^L = \sum_i t_{ij}(\text{pour } a(i)=L) = C_k^L + t_{jk}$, et de même

$$C_j^M = C_k^M - t_{jk} \quad \text{d'où } G_1 + G_2 = 1$$

L'un des deux gains est donc strictement positif.

1.6. *Réduction des données: problème aux images*

Puisque deux objets équivalents sont destinés à être dans la même classe de toute partition centrale il semble naturel de les grouper dès le début et de les traiter comme un seul objet.

Le problème initial, portant sur n objets $i \in E$ est ainsi remplacé par le *problème aux images* portant sur les m images

$$f = a(i) \in a(E) \subset F(n \geqslant m).$$

La similarité s' entre images se définira comme pour les objets et si f et g sont les images de 2 objets i et j quelconques on aura $s_{ij} = s'_{fg} =$ proportion de coordonnées égales de

$$f \text{ et } g \in F_1 \times F_2 \ldots F_p.$$

(Ou plus généralement $s'_{fg} = \sum_1^p p_h s_{fg}^h$.)

Si X est une partition centrale du problème aux objets elle définit une partition X des images puisque deux objets ayant la même image sont dans la même X classe et l'on a $x_{ij} = x'_{fg}$.

Par suite, si n_f et n_g sont les nombres d'objets représentés par f et g dans F on peut les regrouper dans l'expression $L(X)$

$$L(X) = \sum_{i,j} (x_{ij} - s_{ij})^2 = \sum n_f n_g (x'_{fg} - s'_{fg})^2$$

somme étendue à toutes les images f et $g \in a(E)$ (où même à tout $F \times F$).

Les deux expressions en x et x' on les mêmes minima, il en résulte:

Proposition 1.6

Le problème aux m images conduit aux mêmes partitions centrales que le problème aux objets, à condition d'utiliser sur $R^{m \times m}$ la métrique associée à la forme quadratique $Q(X') = \sum n_f n_g (X'_{fg})^2$ (n_f = nombre d'objets d'image f).

1.7. *Métrique du problème aux Images*

La nouvelle métrique utilisée sur $R^{m \times m}$ présente les mêmes qualités algébriques que celles du problème initial.

a) Propriété du barycentre s de masses (p_h, s^h) ($\sum p_h = 1$)

$$\sum p_h Q(x - s^h) = F(x) = Q(x - s) + F(s)$$

b) Le cube des relations $(0,1)^{m \times m}$ est encore inscrit dans la sphère de centre H:

$$h_{fg} = 1/2 \ \forall f, g$$

et d'équation

$$Q(x - h) = \frac{n^2}{4} \ (\text{car} \sum_f n_f = n).$$

c) Par contre la métrique Q n'a pas la même restriction sur le cube des relations $(0,1)^{m \times m}$.

Le carré de la distance entre deux relations x et y, $\sum (x_{ij} - y_{ij})^2$, était le cardinal de la différence symétrique de leurs graphes dans $E \times E$.

Dons le cas présent si x et y sont 2 relations sur F, deux parties de $F \times F$, $Q(x - y)$ représente la somme des produits $n_f n_g$ étendue aux couples f, g qui vérifient une relation et pas l'autre. C'est donc une mesure (à valeurs entières) de la différence symétrique des graphes dans $F \times F$.

Cette mesure est d'ailleurs le produit par elle-même (on pourrait dire "carré cartésien") de la mesure d'ensemble définie sur F par les entiers n_f.

Dans beaucoup de cas particuliers, il pourra être plus naturel de classer les images indépendamment du nombre d'objets qu'elles représentent, et d'utiliser la métrique ordinaire de $R^{m\times m}$. [Cela revient à admettre

$$\begin{aligned} n_f &= 1 \text{ si } f \in a(E) \\ &= 0 \quad f \notin a(E)] \end{aligned}$$

Remarque

La métrique associée à $Q(x)$ est définie par la distance

$$D(x,y) = \sqrt{Q(x-y)}.$$

On peut noter que $Q(x-y) = D^2$ est également une distance sur le cube $(0,1)^{m\times m}$. L'inégalité du triangle:

$$Q(x-z) \leqslant Q(x-y) + Q(y-z)$$

exprime que les triangles inscrits dans ce cube n'ont jamais d'angle obtus. Prenons la métrique ordinaire de $R^{m\times n}$ et le produit scalaire:

$$\vec{xy}.\vec{xz} = \sum_c (x_c - y_c)(x_c - z_c) \quad c = 1, 2, \ldots, n^2$$

x, y et z étant des relations ($x_c, y_c, z_c = 0$ ou 1) chaque terme est positif, car les facteurs sont toujours

$$\begin{aligned} &\geqslant 0 \text{ si } x_c = 1 \\ &\leqslant 0 \text{ si } x_c = 0 \end{aligned}$$

Le cosinus angulaire est donc positif et l'angle yxz aigu. L'angle est droit si pour tout couple c

$$x_c = y_c \text{ ou } z_c$$

soit $\text{Inf}(y_c, z_c) \leqslant x_c \leqslant \text{Sup}(y_c, z_c)$.

En terme de relations:

$$y \text{ et } z \Rightarrow x \Rightarrow y \text{ ou } z,$$

et en terme de graphes

$$y \cup z \subset x \subset y \cap z.$$

Dans l'inégalité du triangle écrite ci-dessus l'égalité se réalise si et seulement et seulement si la relation x est comprise entre la conjonction et la disjonction de y et z.

1.8. *Bonnes Métriques de $R^{n\times n}$ ou $R^{m\times m}$*

On peut se demander quelles sont les métriques de $R^{n\times n}$ qui possèdent les propriétés (a) (b) (c) de la métrique ordinaire.

a) La propriété du barycentre sera vraie pour toute métrique associée à une forme quadratique

$$Q(x) = B(x,x),$$

$B(x,y)$ étant une forme bilinéaire. Plusieurs formes B définissent la même forme Q, mais une seule est symétrique

$$B(x,y) = 1/2[Q(x+y) - Q(x) - Q(y)].$$

b) Par contre seules des formes quadratiques non dégénérées assez particulières rendent toutes les relations $X \in C = (0,1)^{n\times n}$ équidistantes d'un même point G

$$Q(G-X) = \text{cste pour } X \in C$$

(Ce qui permet de réduire le problème quadratique à un problème linéaire).

Car

(1) G est nécessairement confondu avec le point H de coordonnées 1/2. Car H est toujours équidistant d'une relation $X = (x_{ij})$ et de sa négation $X' = (1 - x_{ij})$. Alors si les $2^{n\times n}$ hyperplans médiateurs des couples X, X' avaient un autre point d'intersection G, la droite GH serait "Q-orthogonale" à tous les vecteurs $X - X'$, qui forment un système générateur, et par suite à tout l'espace $R^{n\times n}$. La forme Q serait dégénérée.

(2) Si $Q(X-H) = \text{cste} = R^2$ pour $X \in C$

$Q(X) = 4R^2$ pour $x_{ij} = \pm 1$

le développement

$$Q(X) = \sum_{c,d} q_c^d x_c x_d$$

pour tous les couples c et $d \in E \times E$ ne peut comporter que des termes rectangles

$$c \neq d \Rightarrow q_c^d = 0 \text{ sinon } Q(X)$$

changerait de valeur en remplaçant x_c par $-x_c$.

$Q(X)$ est donc de la forme $\sum q_{ij} x_{ij}^2$.

Si X et Y sont deux relations, $Q(X-Y)$ est ici encore la mesure du cardinal de la différence symétrique de leur graphes, pour la mesure q_{ij} positive sur l'ensemble $E \times E$.

c) Enfin il existe beaucoup de métriques de $R^{n \times n}$ dont la restriction à l'ensemble C des relations soit une métrique donnée, mais une au plus qui dérive d'une forme quadratique $Q(X) = B(X,X)$.

Car les relations b:

$$b_{ij} = 1 \text{ pour 1 seul couple } c = (i,j)$$

forment une base de $R^{n \times n}$ et la somme de deux telles relations est encore une relation. Pour deux vecteurs de base, on a:

$$2B(b,b') = Q(b+b') - Q(b) - Q(b')$$

ce qui définit entièrement la forme bilinéaire symétrique B.

Remarque

Les formes quadratiques associées à la propriété (b) possèdent la propriété des angles aigus déjà vue: $Q(X-Y)$ vérifie l'inégalité du triangle aussi bien que $\sqrt{Q}$.

On démontre aisément que ce sont les seules. Et toutes ces propriétés sont également vraies dans un espace R^H pour H fini.

1.9. *Influence de la Métrique sur le Résultat*

Les métriques Q associées à la propriété b permettent de rendre le problème linéaire:

$$Q(X-S) \text{ minimum} \Leftrightarrow B[HX, HS] \text{ maximum}$$

Cette forme linéaire en X peut s'écrire (à une constante près)

$$L(X) = T'.X = \sum t'_{ij} x_{ij}$$

ou $t'_{ij} = q_{ij}(s_{ij} - 1/2)$.

On voit que la pondération $q_{ij} (> 0)$ détermine l'orientation du vecteur T' dans le cône convexe défini par les contraintes:

$$\text{signe de } t'_{ij} = \text{signe de } s_{ij} - 1/2$$

Quelles que soient les q_{ij}, le maximum de $T'.R$ sur l'ensemble de toutes les relations R est réalisé quand R vérifie:

$$R_{ij} = 1 \quad \text{si } s_{ij} > 1/2$$
$$R_{ij} = 0 \quad \text{si } s_{ij} < 1/2$$

(R_{ij} quelconque pour $s_{ij} = 1/2$)

Cela définit un ensemble de *relations centrales* qui sont indépendantes de la pondération Q sur $E \times E$.

Si l'une d'elles est une relation d'équivalence, elle définit une *partition centrale indépendante de* Q.

Inversement, toute partition centrale X indépendante de Q doit être une relation centrale:

En effet soit A une numérotation des classes de X. Aucun transfert ne doit être avantageux, à partir de A, quel que soit Q. C'est à dire que le transfert d'un objet i, de sa classe L, dans une autre classe M, ne doit jamais être avantageux.

Soit, dans les notations de l'algorithme:

$$C_i^L \geqslant C_i^M, \text{ quels que soient } i, M \text{ et les } q^{ij} > 0$$

$$\text{Comme ici } C_i^M = \sum q_{ij} t_{ij} \text{ pour } A(j) = M$$

cette inégalité implique,

$$t_{ij} = s_{ij} - 1/2 \geqslant 0 \text{ pour } A(j) = L$$
$$\leqslant 0 \text{ pour } A(j) = M$$

Bref: $X_{ij} = 1$ entraîne $s_{ij} \geqslant 1/2$

$= 0$ entraîne $s_{ij} \leqslant 1/2$

X est bien une relation centrale.

Après celà, il est clair qu'en général, les partitions centrales ne sont pas des relations centrales, et dépendent de la pondération Q.

1.10. *Généralisations à d'autres Types de Problèmes*

La notion de relation centrale peut se généraliser sans changement dans toutes les catégories de relations envisageables (ordres, par exemple) Etant données p relations R d'un type T sur un ensemble E fini éventuellement pondérées, on définira le carré de la distance de 2 telles relations comme une mesure positive de la différence symétrique de leur graphe dans $E \times E$.

Une relation X du type T sera dite centrale si la moyenne (éventuellement pondérée) des carrés de ses distances aux p relations R est minimum.

En identifiant T avec une partie du cube $C=(0,1)^{E\times E}$ dans $R^{E\times E}=V$, la métrique étendue à C n'admet qu'une seule extension qui dérive d'une forme quadratique Q, sur V.

Pour cette métrique X sera une projection sur T du barycentre des p relations R, et comme l'ensemble T est inscrit dans une hyper-sphère, le calcul de X se ramène à la recherche du maximum d'une forme linéaire $L(X)$.

Il reste à choisir un algorithme adapté à la nature de l'ensemble T.

ANNEXE

Mise en oeuvre de l'algorithme des transferts

L'algorithme des transferts (1.4.2. p. 181) a été programmé en FORTRAN sur 704 et 1107.

Les paramètres: n = nombre d'objets.

p = borne supérieure du nombre de classes

sont soumis aux contraintes suivantes:

IBM 704 $\quad n \leqslant 100 \quad p \leqslant n$

UNIVAC 1107 $\quad n \leqslant 500 \quad np \leqslant 10.000$

Ces contraintes permettent de maintenir la matrice des attractions (C_i^f = attraction de l'objet i vers la classe f) en mémoire centrale.

Les données proprement dites sont:

La matrice de similarité: $S(n \times n)$

Un classement initial arbitraire: $A(n)$

La sortie correspondante est un classement A', avec la valeur de la fonction économique $L(A')$.

A' est un maximum local: $L(A')$ ne peut augmenter dans aucun transfert.

Le programme 1107 teste en plus si $L(A')$ peut augmenter par réunion de deux classes, ou inversement par transfert d'un objet dans une classe vide, en augmentant la borne supérieure p. On peut imprimer les maxima locaux correspondants à des valeurs de p en progression arithmétique: (po, Dp), et obtenir en particulier (pour $Dp=1$) l'un des maximum locaux qui comporte le plus petit nombre de classes.

Quant à garantir un maximum comme absolu, la question théorique reste ouverte. Nous ne savons même pas borner supérieurement le nombre $l=1$ de maxima locaux ni prévoir si $l=1$. (!). Une précaution élémentaire consiste à faire 2 calculs en prenant comme classement initial:

$A(I) = 1$ (n classes d'1 objet chacune)

$A(I) = 1$ 1 seule classe.

Dans de nombreux cas, nous avons obtenu deux fois le même maximum local A'. On peut alors présumer que A' est un maximum général. Les temps de calcul étaient de l'ordre de:

12 minutes sur 704

3 minutes sur 1107

pour un seul calcul $A \rightarrow A^{l}$.

REMERCIEMENTS

L'auteur tient à remercier spécialement:

M. Hans, de l'Institut Blaise Pascal du C.N.R.S. qui a bien voulu s'associer à ce travail et programmer et tester l'algorithme des transferts sur IBM 704.

Mme. Renaud, de la Section d'Automatique Documentaire du C.N.R.S. qui a effectué l'extension de ce programme sur l'UNIVAC 1107 de la Faculté des Sciences d'Orsay.

Enfin, M. Bernard Jaulin, Directeur du Centre de Calcul de la Maison des Sciences de l'Homme, qui s'est intéressé à ce travail d'un bout à l'autre et m'a souvent aidé de ses suggestions constructives.